CLAIMS:

1. A system for controlling and managing Internet server farm traffic through a plurality of servers, the server farm traffic arriving at a server farm as inbound traffic organized by customer (i) and traffic type (j) and leaving the server farm as outbound traffic, the system being operable to control and manage the outbound traffic in accordance with outbound bandwidth usage-based service level agreements of form (Bmin,Bmax), the system comprising:

means for collecting the admitted rate (Ra) of inbound traffic for each customer traffic type (i,j);

means for collecting the rejected rate (Rr) of inbound traffic for each customer traffic type (i,j);

means for collecting the outbound traffic (B) for each customer traffic type (i,j);

means for computing an expected bandwidth usage (b) per TCP connection request for each customer traffic type (i,j);

means for computing the target rate (Rt) for each customer traffic type (i,j) that supports the outbound bandwidth usage-based service level agreements of form (Bmin,Bmax);

limiter means for admitting inbound traffic based on the target rate (Rt) and for tracking the volume of admitted inbound traffic (Ra) and the volume of rejected inbound traffic (Rr) for each customer traffic type (i,j);

means for relaying the target rates (Rt) for inbound traffic to the limiter means; and

means for dispatching the admitted inbound traffic (Ra) to the servers.

2. A system according to claim 1, wherein the means for collecting the admitted rate (Ra) and the rejected rate (Rr) of inbound traffic comprises an inbound traffic scheduler device and an inbound traffic monitor, the inbound traffic monitor

being operable to observe the admitted rate (Ra) and the rejected rate (Rr) of inbound traffic and relay the admitted rate (Ra) and rejected rate (Rr) to the inbound traffic scheduler device.

3. A system according to claim 2, wherein the inbound traffic monitor is associated with the dispatching means.

4. A system according to claim 1, wherein the means for collecting the admitted rate (Ra) and the rejected rate (Rr) of inbound traffic comprises an inbound traffic scheduler device and the limiter means, the limiter means being operable to observe and relay the amount of admitted inbound traffic (Rp) and the amount of rejected traffic (Rr) to the inbound traffic scheduler device.

5. A system according to claim 4, wherein the limiter means is associated with the dispatching means.

6. A system according to claim 1, wherein the means for collecting the outbound traffic (B) comprises an inbound traffic scheduler device and an outbound traffic monitor, the outbound traffic monitor being operable to observe and relay the amount of outbound traffic (B) to the inbound traffic scheduler device.

7. A system according to claim 6, wherein the outbound traffic monitor is associated with the servers.

8. A system according to claim 1, further comprising means for observing the average resource usage (c) of each server consumed for each consumer traffic type (i,j).

9. A system according to claim 8, wherein the means for observing the

average resource usage (c) is associated with the servers.

10. A system according to claim 1, further comprising means for dispatching the inbound traffic among the servers.

11. A system according to claim 1, wherein the dispatching means comprises at least one inbound traffic limiter, a high-speed LAN and a plurality of dispatchers, the limiter means being associated with the inbound traffic limiter, each of the dispatchers being associated with at least one of the servers.

12. A system according to claim 1, wherein the dispatching means comprises a high-speed LAN and a plurality of dispatchers, the limiter means and monitor means being associated with each of the dispatchers, each of the dispatchers being associated with at least one of the servers.

13. A system according to claim 1, further comprising means for establishing an absolute bound (Rbound) of the target rate (Rt) for each customer traffic type (i,j).

14. A system according to claim 13, further comprising means for collecting the absolute bound (Rbound) of the target rate (Rt) for each customer traffic type (i,j).

15. A system according to claim 14, further comprising means for limiting the target rate (Rt) for inbound traffic when Rbound is available from the establishing means.

16. A system for controlling and managing Internet server farm traffic through a plurality of servers, the server farm traffic arriving at a server farm as

inbound traffic organized by customer (i) and traffic type (j) and leaving the server farm as outbound traffic, the system being operable to control and manage the outbound traffic in accordance with outbound bandwidth usage-based service level agreements of form (Bmin,Bmax) and in accordance with a server resource manager that establishes an absolute bound (Rbound) of a target rate (Rt) for each customer traffic type (i,j), the system comprising:

an inbound traffic scheduler device operable to collect the admitted rate (Ra) and the rejected rata (Rr) of inbound traffic for each customer traffic type (i,j), collect the bound (Rbound) from any server resource manager and collect the outbound traffic(B) for each customer traffic type (i,j), the inbound traffic scheduler device being further operable to compute an expected bandwidth usage (b) per request for each customer traffic type (i,j) and compute a target rate (Rt) for inbound traffic for each customer traffic type (i,j) to support the outbound bandwidth usage-based service level agreements of form (Bmin,Bmax);

an inbound traffic limiter operable to receive the target rate (Rt) from the inbound traffic scheduler device, admit inbound traffic based on the target rate (Rt), track the volume of admitted inbound traffic (Ra) and the volume of rejected inbound traffic (Rr) for each customer traffic type (i,j), and relay the amount of admitted inbound traffic (Ra) and the amount of rejected traffic (Rr) to the inbound traffic scheduler device; and

an inbound traffic dispatching network operable to classify incoming traffic, the inbound traffic dispatching network being controlled by the inbound traffic limiter to selectively dropping packets arriving in the inbound traffic limiter, the inbound traffic dispatching network further being comprised of a high-speed LAN with dispatchers to dispatch the admitted inbound traffic (Ra) to the servers.

17. A system according to claim 16, wherein the inbound traffic scheduler is operable to compute target rates (Rt) for all customer traffic type (i,j) to meet with the service level agreements of form (Bmin,Bmax) on the outbound

bandwidth usage, and is operable to support both bandwidth borrowing and bandwidth not-borrowing modes of operations.

18. A system according to claim 16, wherein the inbound traffic limiter is associated with the inbound traffic dispatching network.

19. A system according to claim 16, further comprising an inbound traffic monitor associated with the inbound traffic dispatching network, the inbound traffic monitor being operable to observe the admitted rate (Ra) and the rejected rata (Rr) of inbound traffic and relay the admitted rate (Ra) and the rejected rate (Rr) to the inbound traffic scheduler device.

20. A system according to claim 16, further comprising an outbound traffic monitor that is operable to observe and relay the amount of outbound traffic (B) to the inbound traffic scheduler device.

21. A system according to claim 16, further comprising a resource usage monitor that is operable to observe and relay the average resource usage (c) of each server consumed for each consumer traffic type (i,j) to the inbound traffic scheduler device.

22. A system according to claim 16, wherein the inbound traffic dispatching network is operable to balance the inbound traffic among the servers.

23. A system according to claim 16, wherein the inbound traffic dispatching network comprises at least one inbound traffic limiter, a high-speed LAN and a plurality of dispatchers, each of the dispatchers being associated with at least one of the servers.

24. A system according to claim 16, wherein the inbound traffic dispatching network comprises a high-speed LAN and a plurality of dispatchers, the inbound traffic limiter and inbound traffic monitor being associated with each of the dispatchers, each of the dispatchers being associated with at least one of the servers.

25. A method for controlling and managing Internet server farm traffic through a plurality of servers, the server farm traffic arriving at a server farm as inbound traffic organized by customer (i) and traffic type (j) and leaving the server farm as outbound traffic that is controlled and managed in accordance with outbound bandwidth usage-based service level agreements (Bmin(i,j),Bmax(i,j)), the method comprising the steps of:

collecting the admitted rate (Ra(i,j)) of inbound traffic for each customer traffic type (i,j);

collecting the rejected rate (Rr(i,j)) of inbound traffic for each customer traffic type (i,j);

collecting the outbound traffic (B(i,j)) for each customer traffic type (i,j);

collecting the absolute bound (Rbound(i,j)) on the target rate (Rt(i,j)) for each customer traffic type (i,j);

computing an expected bandwidth usage (b(i,j)) per TCP connection request for each customer traffic type (i,j);

computing the target rate (Rt(i,j)) for each customer traffic type (i,j) based on the admitted rate (Ra(i,j)), the rejected rate (Rr(i,j)), the outbound traffic (B(i,j)), the expected bandwidth usage (b(i,j)) and the outbound bandwidth usage-based service level agreements (Bmin(i,j),Bmax(i,j));

admitting inbound traffic based on the target rate (Rt(i,j)) and tracking the volume of admitted inbound traffic (Ra(i,j)) and the volume of rejected inbound traffic (Rr(i,j)) for each customer traffic type (i,j);

relaying the target rates (Rt(i,j)) for inbound traffic to the limiter

means; and

dispatching the admitted inbound traffic (Rp(i,j)) to the servers.

26. A method according to claim 25, wherein the step of collecting the admitted rate (Ra(i,j)) and the rejected rate (Rr(i,j)) of inbound traffic comprises the steps of observing the admitted rate (Ra(i,j)) and the rejected rate (Rr(i,j)) of inbound traffic with an inbound traffic monitor and then relaying the admitted rate (Ra(i,j)) and the rejected rate (Rr(i,j)) to an inbound traffic scheduler device that performs the steps of computing the expected bandwidth usage (b(i,j)) and the target rate (Rt(i,j)) to support the outbound bandwidth usage-based service level agreements (Bmin(i,j),Bmax(i,j)).

27. A method according to claim 25, wherein the step of collecting the admitted rate (Ra(i,j)) and the rejected rate (Rr(i,j)) of inbound traffic comprises the steps of observing the amount of admitted inbound traffic (Ra(i,j)) and the amount of rejected traffic (Rr(i,j)) with an inbound traffic limiter and then relaying the admitted rate (Ra(i,j)) and the rejected rate (Rr(i,j)) to an inbound traffic scheduler device that performs the steps of computing the expected bandwidth usage (b(i,j)) and the target rate (Rt(i,j)) to support the outbound bandwidth usage-based service level agreements (Bmin(i,j),Bmax(i,j).

28. A method according to claim 25, wherein the step of collecting the bound (Rbound(i,j)) on the target rate (Rt(i,j)) comprises the steps of receiving the bound (Rbound(i,j)) from a server resource manager.

29. A method according to claim 25, wherein the step of collecting the outbound traffic(B(i,j)) comprises the steps of observing the amount of outbound traffic (B(i,j)) with an outbound traffic monitor and then relaying the amount of outbound traffic(B(i,j)) to a device that performs the steps of computing the expected

bandwidth usage (b(i,j)) and the target rate (Rt(i,j)) to support the outbound bandwidth usage-based service level agreements (Bmin(i,j),Bmax(i,j)).

30. A method according to claim 25, further comprising the step of observing the average resource usage (c(i,j)) of each server consumed for each consumer traffic type (i,j).

31. A method according to claim 25, further comprising the step of balancing the inbound traffic among the servers.

32. A method according to claim 25, further comprising the step of limiting the target rate (Rt(i,j)) for inbound traffic independently of the service level agreement (Bmin(i,j), Bmax(i,j)).

33. A method according to claim 25, further comprising the step of classifying incoming traffic and selectively dropping packets prior to admitting and dispatching the packets to the servers.